CRYPTO CURRENCY PRICE PREDICTOR

GROUP MEMBERS

Muzammil Rizvi (CS-21089)

Muhammad Ali Hasnain (CS-21068)

Muhammad Ashhad Jamal (CS-21098)

Abdul Kareem Ahmed (CS-21102)

JULY 8, 2024

GROUP: G3

CIS DEPARTEMENT

CRYPTO-CURRENCY PRICE PREDICTOR PROJECT REPORT

INTRODUCTION:

This project conducted a comprehensive analysis and prediction of cryptocurrency prices for Dogecoin, Ethereum, Cardano, and Tether. Each coin's dataset was processed through a pipeline consisting of Exploratory Data Analysis (EDA), preprocessing, feature engineering, model selection, model training, and evaluation. The objective was to develop accurate predictive models using various regression techniques and compare their performance across different cryptocurrencies.

DATASET:

The datasets for Dogecoin, Ethereum, Cardano, and Tether were obtained from a reputable cryptocurrency database. Each dataset included features such as the date, open price, high price, low price, close price, volume, and market capitalization. These datasets provided a rich basis for analysis and model training, spanning several years and capturing the dynamic nature of cryptocurrency markets. The datasets allowed us to explore historical trends, identify patterns, and create predictive models with a solid foundation of data.

EXPLORATORY DATA ANALYSIS (EDA):

EDA was performed to understand the distribution and relationships within the data:

- **Dogecoin**: Significant price volatility, especially during certain periods. The volume showed spikes correlating with price surges.
- **Ethereum:** High market capitalization and volume, with clear upward trends over time. The price exhibited periodic peaks corresponding to major market events.
- **Cardano:** Steady growth with occasional spikes. The volume and price showed strong correlation, indicating market interest.
- **Tether:** Stable price due to its nature as a stablecoin, with high volume. The low-price volatility and high trading volume reflected its use as a liquidity provider in the market.

PREPROCESSING:

- **Handling missing values:** We addressed missing data points by filling or interpolating values to ensure data continuity.
- **Converting date formats:** Dates were standardized to a uniform format for consistent analysis across all datasets.
- **Normalizing numerical features:** Price and volume data were normalized to bring all features to a comparable scale, aiding in model convergence.

- **Encoding categorical features:** Although limited, any categorical features were encoded appropriately to be used in the models.

FEATURE ENGINEERING:

Feature engineering involved creating new features to enhance model performance:

- **Dogecoin:** Lag features for previous prices & volumesand moving averages.
- **Ethereum:** Technical indicators like RSI, moving averages, and lag features.
- **Cardano:** Rolling averages, lag features, and market sentiment indicators.
- **Tether:** Lag features, rolling averages, and volume trend indicators.

MODEL SELECTION:

Five regression models were selected for comparison:

1. **Linear Regression:** A simple model that assumes a linear relationship between the features and the target variable.
2. **Random Forest Regressor:** An ensemble learning method that constructs multiple decision trees for better accuracy and robustness.
3. **Gradient Boosting Regressor:** Another ensemble method that builds models sequentially, each correcting the errors of its predecessor.
4. **Support Vector Regressor (SVR):** A model that uses support vectors to perform regression in a high-dimensional space.
5. **Decision Tree Regressor:** A model that splits the data into subsets based on feature values, creating a tree-like structure.

These models were chosen to cover a range of simple to complex techniques, providing a broad comparison for our predictions.

MODEL TRAINING:

Each model was trained on the respective coin's training dataset. Hyperparameters were optimized using cross-validation to ensure the best performance. We trained five models for each of the four coins, resulting in a total of twenty trained models. The training process involved splitting the data into training and testing sets, tuning the models, and ensuring they generalized well to unseen data.

EVALUATION:

Models were evaluated using Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2) metrics. MSE measures the average squared difference between predicted and actual values, MAE measures the average absolute difference, and R^2 indicates the proportion of variance explained by the model. These metrics provided a comprehensive view of model performance.

RESULTS:

Overall, the Linear Regression model provided the best results across all four coins, demonstrating the highest R^2 and lowest MSE and MAE. The Random Forest and Gradient Boosting models also performed well but were slightly less accurate. The SVR and Decision Tree models had more variability in their results. Despite the differences, the models generally showed consistent performance across different cryptocurrencies.

MODEL LIMITATIONS:

Despite their effectiveness, our models have limitations. They heavily rely on historical data and may not anticipate abrupt market changes or external events. Cryptocurrency market volatility poses a challenge to their predictive accuracy. Additionally, the models' performance can vary across different cryptocurrencies due to unique market dynamics and investor behaviours.

WEB APPLICATION:

We developed a web application to showcase our predictions, visualizations, and analysis for each coin. The website features interactive charts, allowing users to explore the historical trends and predicted prices of Dogecoin, Ethereum, Cardano, and Tether. Users can also view detailed analysis and performance metrics for each model, providing a comprehensive understanding of our findings. The application serves as a valuable tool for both educational and practical purposes, demonstrating the power of data-driven predictions in the cryptocurrency market.

FUTURE EXPANSION:

Future enhancements could include integrating sentiment analysis from social media and news sources to better capture market sentiment. Exploring advanced deep learning models like LSTM networks could improve predictions by capturing intricate temporal dependencies. Ensemble techniques tailored for cryptocurrency markets could also enhance performance by combining model strengths. Furthermore, expanding the feature set to encompass blockchain-specific metrics and economic indicators could provide a more comprehensive market analysis.

CONCLUSION:

This project successfully demonstrated the application of various regression techniques to predict cryptocurrency prices. While Linear Regression provided the best results across most coins, different models excelled for specific datasets, highlighting the importance of model selection based on dataset characteristics. The web application offers an accessible way to explore and understand the predictions and analysis. Future work could involve integrating more sophisticated deep learning models and expanding the feature set for even more accurate predictions.

DEPARTMENT OF COMPUTER & INFORMATION SYSTEMS ENGINEERING
BACHELORS IN COMPUTER SYSTEMS ENGINEERING
Course Code: CS-324
Course Title: Machine Learning
Open Ended Lab
TE Batch 2021, Spring Semester 2024
Grading Rubric
TERM PROJECT

Group Members:

Student No.	Name	Roll No.
S1	Muzammil Rizvi	CS-21089
S2	Muhammad Ali Hasnain	CS-21068
S3	Muhammad Ashhad Jamal	CS-21098
S4	Abdul Kareem Ahmed	CS-21102

CRITERIA AND SCALES				Marks Obtained		
				S1	**S2**	**S3**
Criterion1: Data Collection						
0	1	2	3			
The student has not chosen a suitable dataset for predictive modeling.	The student has chosen a dataset, but it may not be suitable for predictive modeling, or it lacks enough features.	The student has chosen a suitable dataset for predictive modeling, and it has enough features to work with.	The student has chosen an excellent dataset for predictive modeling, which has rich features and is well-suited for the task.			
Criterion 2: Data Preprocessing						
0	1	2	3			
The student has not performed data cleaning, handling missing values, or encoding categorical variables	The student has performed basic data cleaning and handled missing values, but has not encoded categorical variables.	The student has performed data cleaning, handled missing values, and encoded categorical variables.	The student has performed thorough data cleaning, handled missing values effectively, and encoded categorical variables efficiently.			
Criterion 3: Exploratory Data Analysis (EDA)						
0	1	2	3			
The student has not performed exploratory data analysis (EDA) or provided minimal analysis with no meaningful insights.	The student has performed basic exploratory data analysis, but the analysis lacks depth, and insights are limited	The student has performed thorough exploratory data analysis, identifying important variables, correlations, and providing meaningful insights.	The student has performed exceptional exploratory data analysis, providing comprehensive insights, and utilizing a variety of visualization techniques effectively.			
Criterion 4: Feature Engineering						
0	1	2	3			
The student has not performed feature engineering.	The student has performed basic feature engineering, but has not created new features or scaled/normalized existing features.	The student has performed feature engineering, creating new features and scaling/normalizing existing features if required.	The student has performed advanced feature engineering, creating meaningful new features and effectively scaling/normalizing existing features.			
Criterion 5: Model Building						
0	1	2	3			
The student has not built any predictive models.	The student has built models using machine learning algorithms, but the implementation lacks depth, and multiple algorithms were not used.	The student has built models using multiple machine learning algorithms, implementing them using Python packages, and evaluated their performance.	The student has built models using multiple machine learning algorithms, implemented them both using Python packages and without Python packages, and			

			thoroughly evaluated their performance.			
Criterion 6: Model Evaluation						
0	1	2	3			
The student has not evaluated model performance or has done so inadequately.	The student has evaluated model performance but has not used different techniques or compared the performance of different models.	The student has evaluated model performance using different techniques, compared the performance of different models, and selected the best-performing model.	The student has thoroughly evaluated model performance using various techniques, performed a detailed comparison of different models, and selected the best-performing model based on comprehensive evaluation metrics.			
Criterion 7: Conclusion						
0	1	2	3			
The student has not provided a conclusion or has provided a conclusion with minimal insights.	The student has provided a basic conclusion with some insights but has not discussed model limitations or suggested improvements.	The student has provided a detailed conclusion with meaningful insights, discussed model limitations, and suggested improvements.	The student has provided an exceptional conclusion with comprehensive insights, thorough discussion of model limitations, and insightful suggestions for improvements.			
Criterion 8: Report						
0	1	2	3			
The submitted report is unfit to be graded.	The report is partially acceptable.	The report is complete and concise.	The report is exceptionally written.			
			Total Marks:			